October 4, 2006

Higher Order Statistical Method for Extracting Dependencies in Multivariate Geospace Data Sets

Jay R. Johnson, Princeton University, Plasma Physics Laboratory
Simon Wing, Johns Hopkins University, Applied Physics Laboratory
Tomaš Gedeon, Montana State University

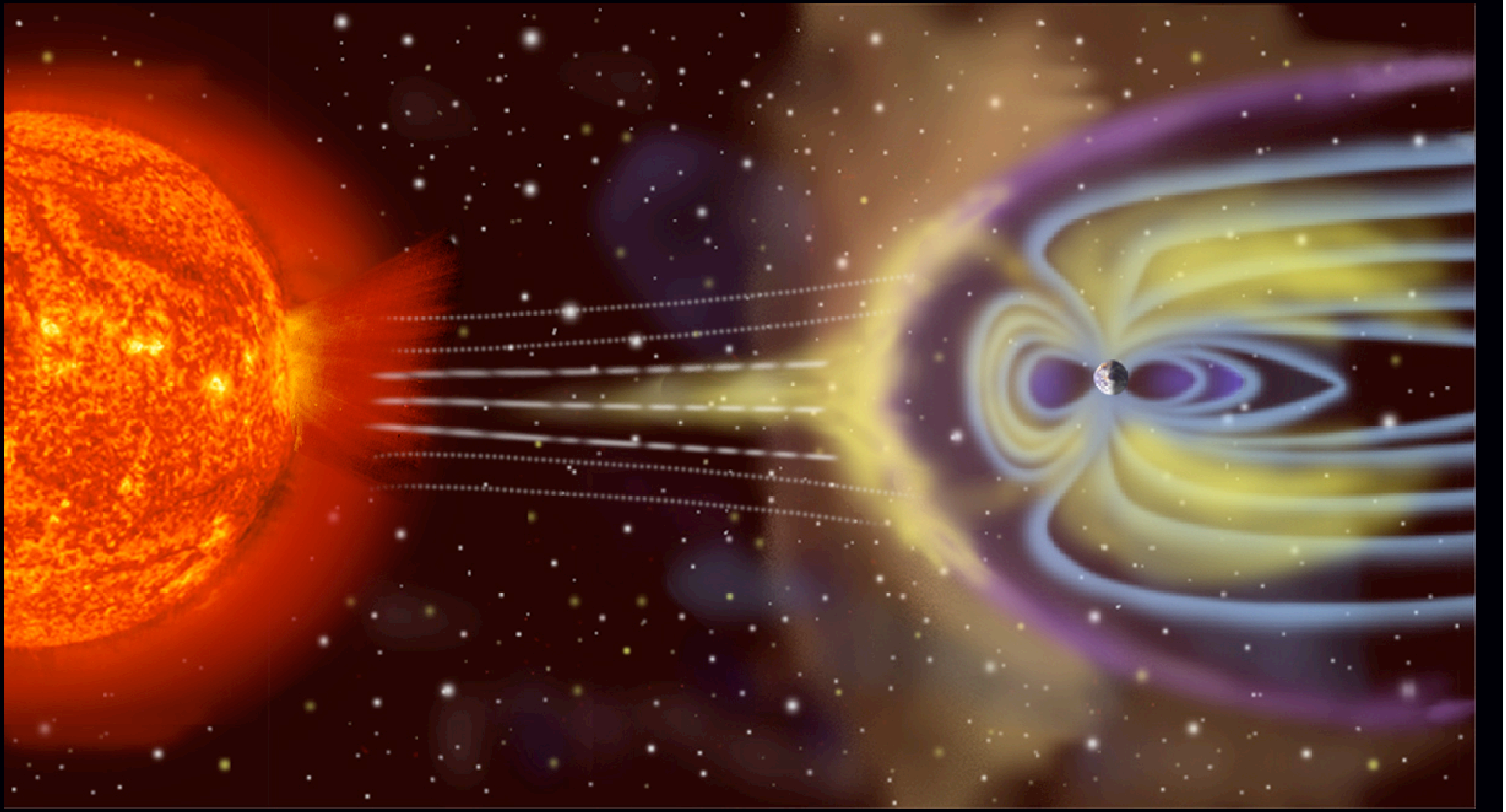
AISRP Investigators' Workshop, College Park, MD

PPPL
PRINCETON PLASMA PHYSICS LABORATORY

Outline

- Objective
- Data Sets
- Methodology
- Planned Activities

Understanding the Evolution of the Magnetospheric State



Objective to Understand

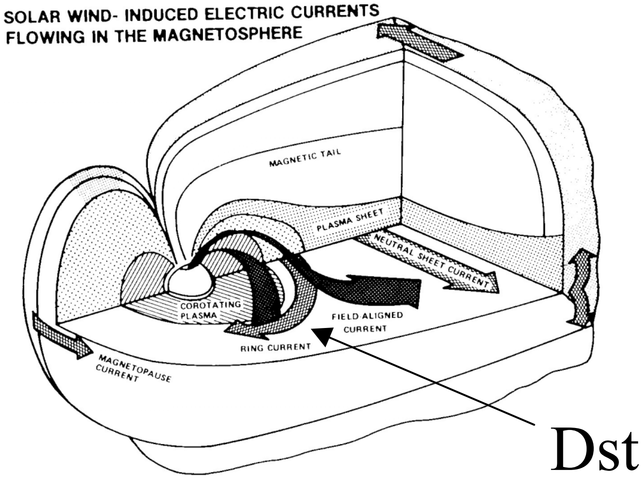
- Underlying dynamics of geospace systems
  - Driven vs Internal Response
  - Dimensionality
- Nonlinear behavior
  - Loading/unloading processes
- Nonstationarity
  - Short term: e.g. substorms
  - Long term: solar cycle changes
- Predictability
  - Horizon
  - Models

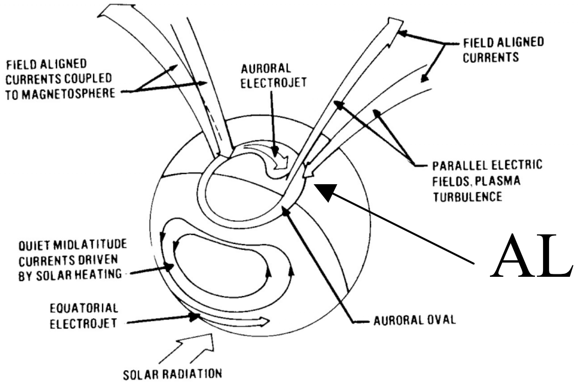
Methodology

- Consider the evolution of variables that define the "state" of the magnetosphere:
  - Geomagentic indices: Kp, Dst, AL, …
  - Physical Measures: Particle flux, b2i, auroral power
- Determine the dependence on:
  - External solar wind drivers: n,V,B,P,…
  - History of the magentospheric state

Defining the State of the Magnetosphere

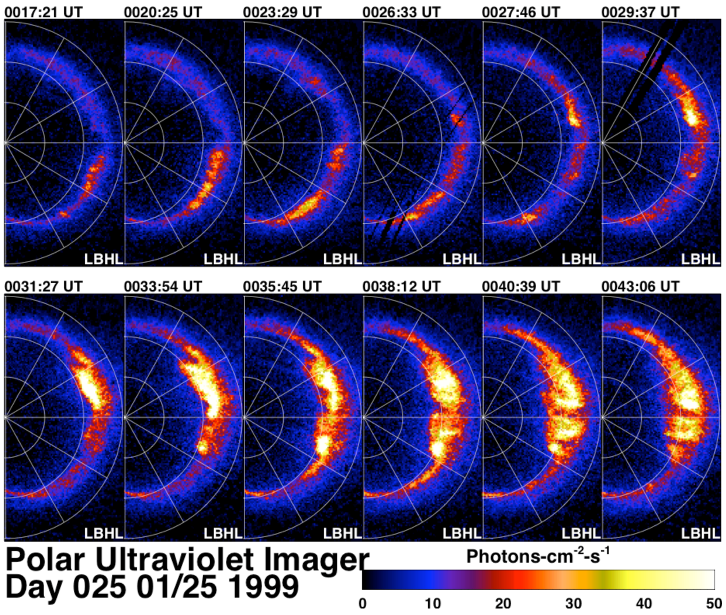
- Magnetic indices
- Historical record (1932 to present)
- Related to physical current distributions

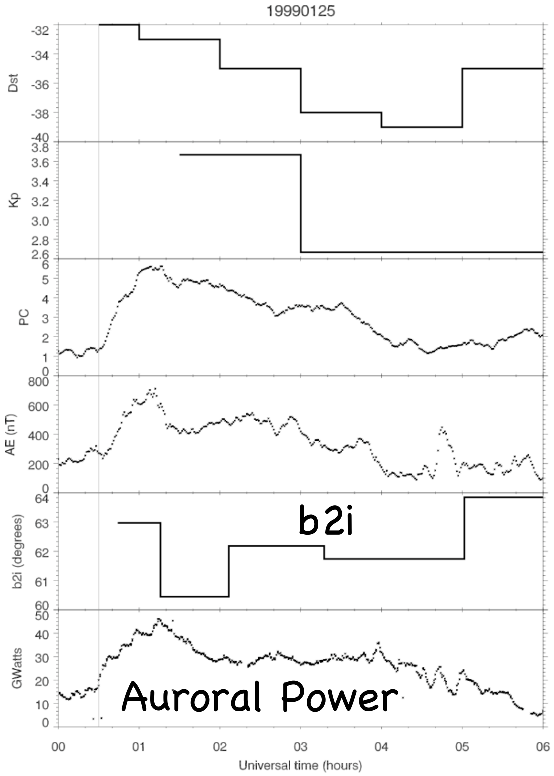



FIELD ALIGNED CURRENTS COUPLED TO MAGNETOSPHERE
AURORAL ELECTROJET
FIELD ALIGNED CURRENTS
PARALLEL ELECTRIC FIELDS, PLASMA TURBULENCE
AL
QUIET MIDLATITUDE CURRENTS DRIVEN BY SOLAR HEATING
EQUATORIAL ELECTROJET
AURORAL OVAL
SOLAR RADIATION

Defining the State of the Magnetosphere

- auroral power
- polar cap potential
- b2i---tail stretching
- energetic particle flux

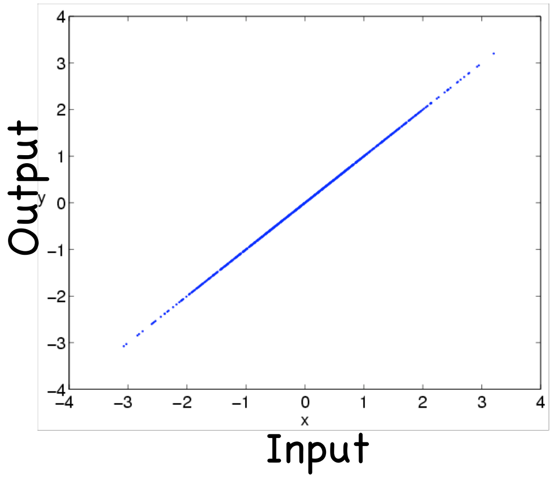


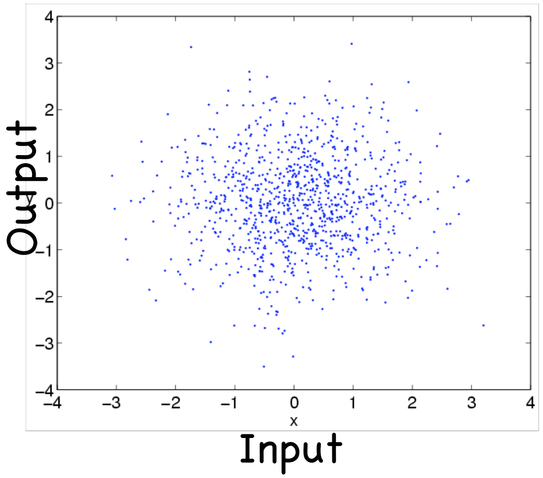


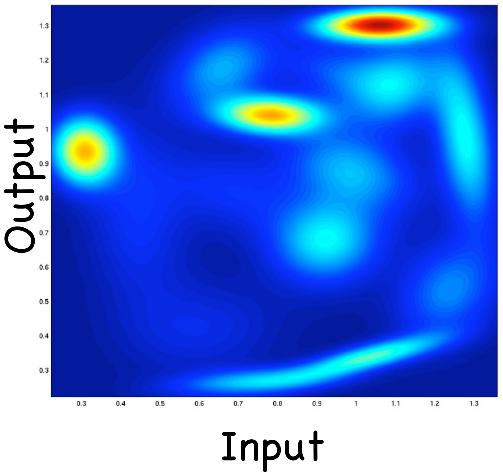
Statistical Measures of Dependency

1. Correlation function
2. Mutual Information
3. Cumulants

$$P(x,y) = P(x)P(y)\ ?$$


Correlated Data


Uncorrelated Data


High Order Correlations

Entropy and Mutual Information

$$H(x) = -\sum_{\aleph_1} p(x)\log p(x)$$

$$H(y) = -\sum_{\aleph_2} p(y)\log p(y)$$

$$H(x,y) = -\sum_{\aleph_1,\aleph_2} p(x,y)\log p(x,y)$$

$$I(x,y) = H(x) + H(y) - H(x,y)$$

$$x \in \{1,\ldots,N\} \equiv \aleph_1$$

$$y \in \{1,\ldots,M\} \equiv \aleph_2$$

Mutual information is commonly used as an alternative to correlation functions which have limitations for nonlinear systems. Generalization to higher dimensions is called redundancy.

Discriminating Statistic
Mutual Information

$$\lambda(\mathbf{X},\mathbf{Y}) \equiv \sqrt{1 - \frac{\det C(\mathbf{X},\mathbf{Y})}{\det C(\mathbf{X}) \det C(\mathbf{Y})}}$$

$$\Lambda(\mathbf{X},\mathbf{Y}) \equiv \sqrt{1 - \exp(-2I(\mathbf{X},\mathbf{Y}))}$$

$D_{MI} = \Lambda - \lambda$

$\Lambda \Im \lambda$

when Gaussian distributed joint PDF

Discriminating Statistic
Cumulant Based Cost

P(x,y) = P(x)P(y) ?

Statistical Independence ⇒ Cumulants vanish

D_C= measure of cross cumulants

$$D = \sum_{n=1}^{\infty} \sum_{i_2 \ldots i_n = 1}^{m} (1 - \delta_{1 i_2 \ldots i_n}) \{K_{1 i_2 \ldots i_n}\}^2$$

Multivariate Cumulants

$$C_{i,\ldots,j} = \left\langle x_i \cdots x_j \right\rangle$$

← Correlation Tensors

$$K_i = C_i = \left\langle x_i \right\rangle$$

Cumulants

↓

$$K_{i,j} = C_{i,j} - C_i C_j = \left\langle x_i x_j \right\rangle - \left\langle x_i \right\rangle \left\langle x_j \right\rangle$$

$$K_{i,j,k} = C_{i,j,k} - C_{i,j} C_k - C_{j,k} C_i - C_{i,k} C_j + 2\, C_i C_j C_k$$

$$K_{i,j,k,l} = C_{i,j,k,l} - C_{i,j,k} C_l - C_{i,j,l} C_k - C_{i,l,k} C_j - C_{l,j,k} C_i$$

$$- C_{i,j} C_{k,l} - C_{i,l} C_{k,j} - C_{i,k} C_{j,l}$$

$$+ 2\, \big(C_{i,j} C_k C_l + C_{i,k} C_j C_l + C_{i,l} C_j C_k + C_{j,k} C_i C_l + C_{j,l} C_i C_k + C_{k,l} C_i C_j \big)$$

$$- 6\, C_i C_j C_k C_l$$

Discriminating Statistics Detect Nonlinearity

- D_{MI} measures deviance from linear correlations
- Truncating D_C at second order vs higher order provides information about nonlinear dependency
- Practical Point: D_C gives better statistics for limited and noisy data

How can we know
if these measures have
any meaning?

Comparison with Surrogate Data

$$S = \frac{|\, D_0 - \mu_s \,|}{\sigma_s}$$

Significance Measured
Relative to a Null Hypothesis

$$\mu_s = \frac{1}{N}\sum_{i=1}^{N} D_{S_i}$$

← Surrogates Generated by
CAAFT process

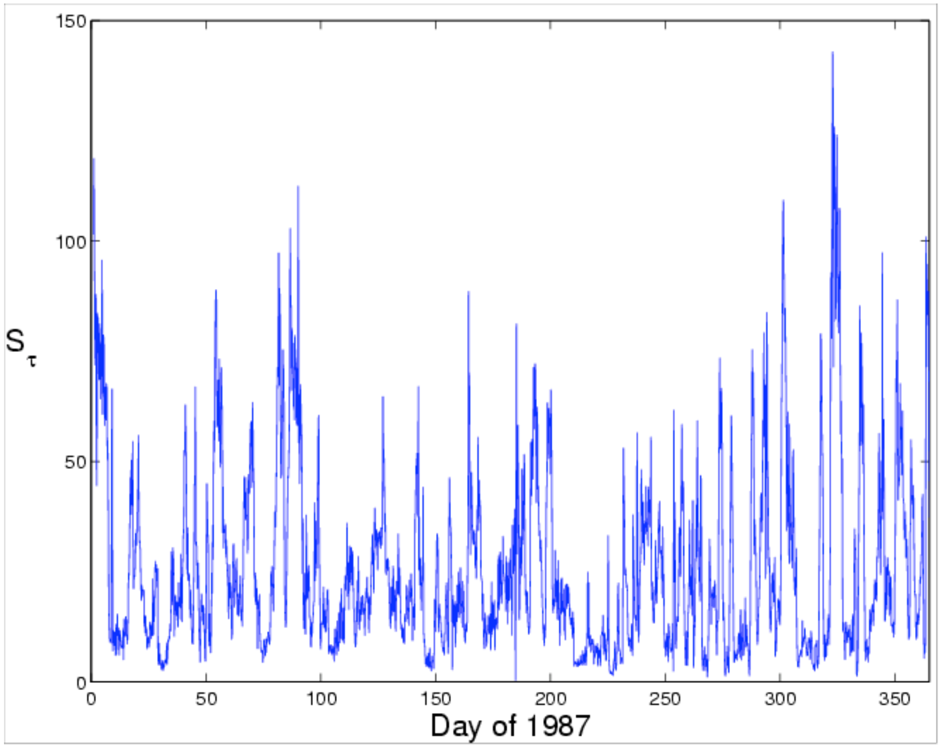
$$\sigma_s^{\,2} = \frac{1}{N-1}\sum_{i=1}^{N}(D_{S_i} - \mu_s)^2$$

An Example: the underlying dynamics of Kp

- Obtain nonlinear predictability measure (Λ - λ) and cumulant-based cost, D, for the original data
- Construct surrogate data having the same linear properties as the original data (λ)
  - CAAFT method
- Evaluate nonlinear predictability and cumulant-based cost for the surrogate data
- Use the significance test to determine when the null hypothesis (linear dynamics) is invalid

Nonstationarity Indicates Multiple Magnetospheric Dynamical Modes

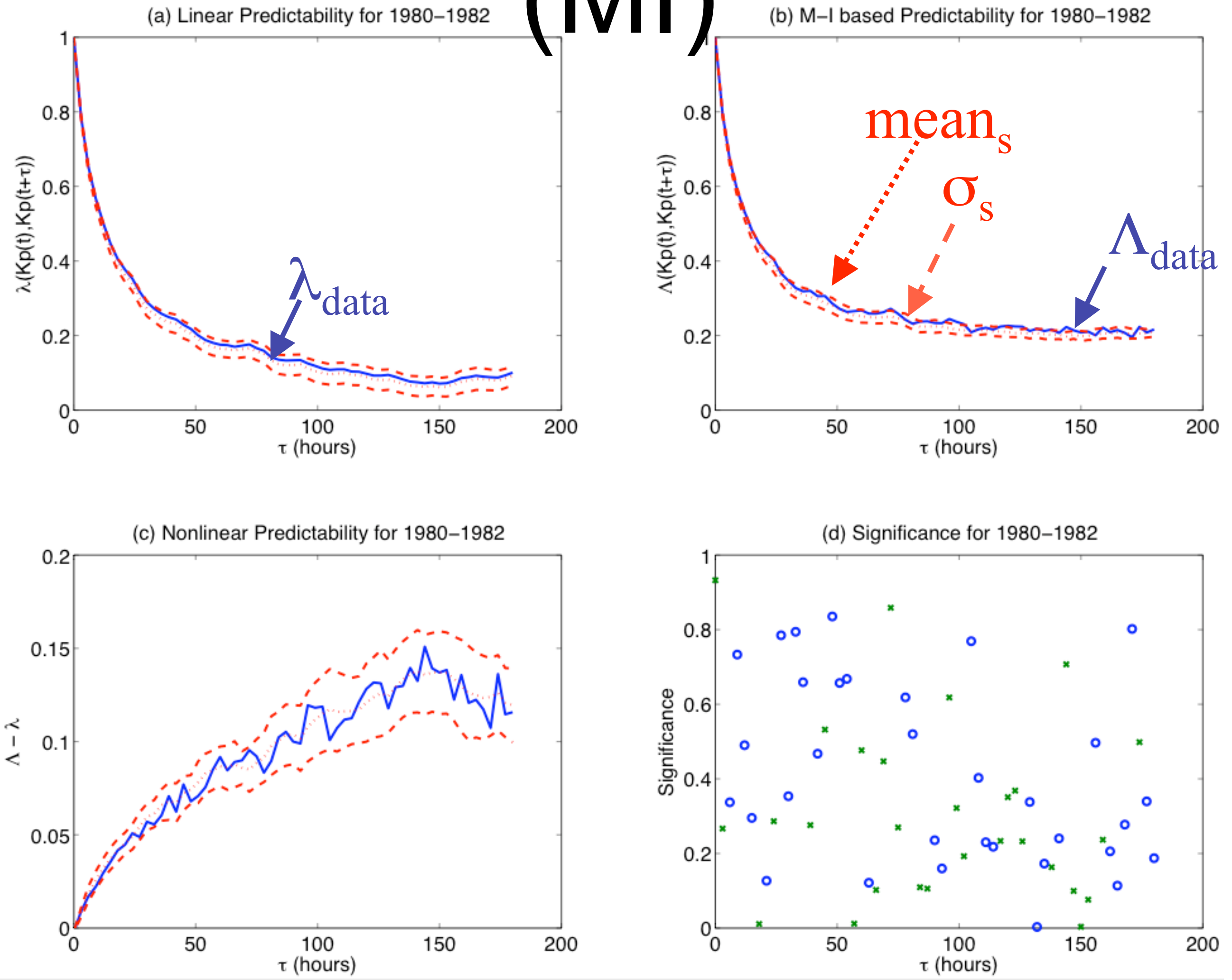
Variation in Underlying Magnetospheric Dynamics



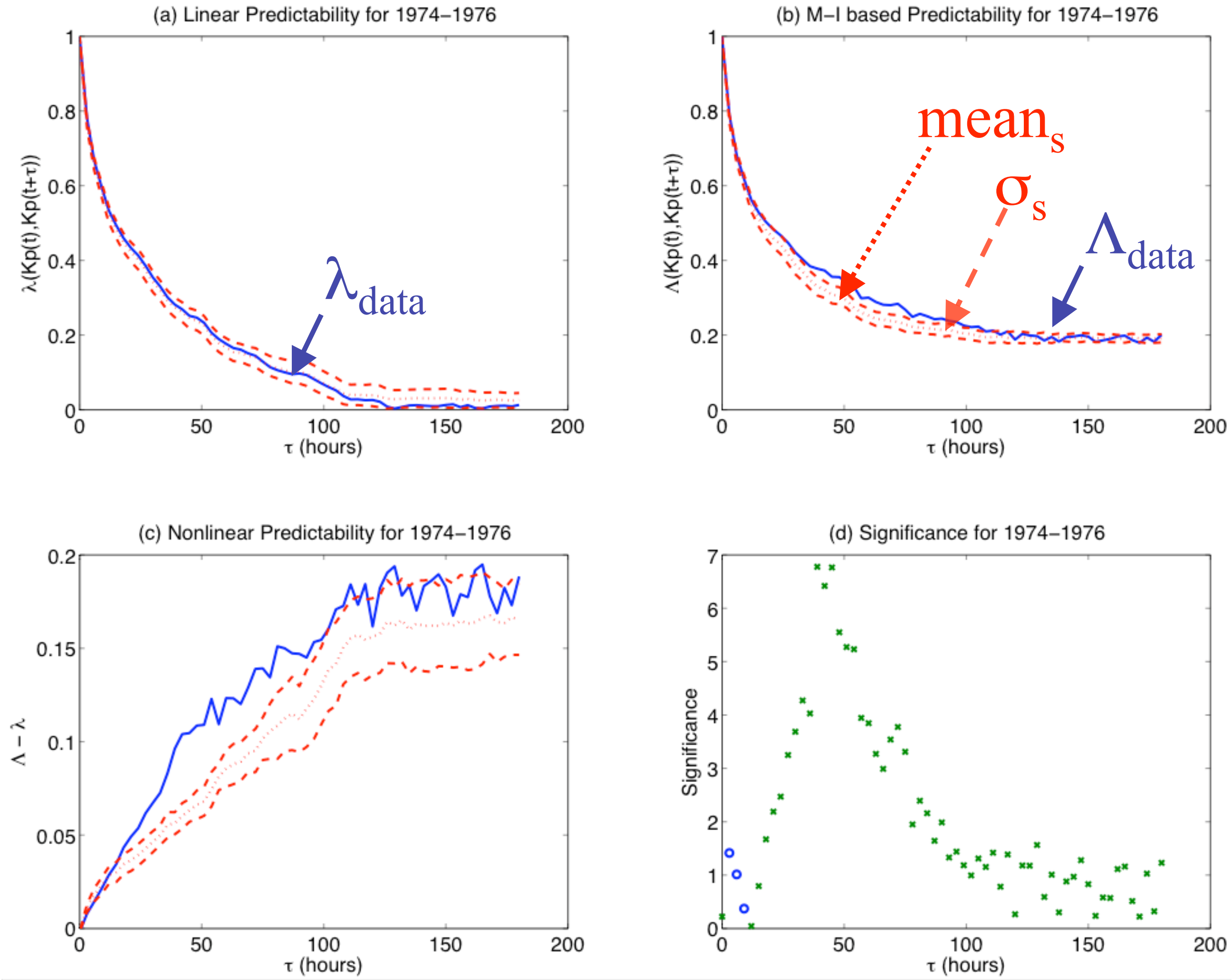
- $m=3, \Delta=1\text{hr}, N_W=300\text{hrs}$
  $N_S=100$
- $S >> 1$ means strong time ordering and good predictability over ~10 days
- Large Variations in S imply changes in underlying dynamics
- Several dynamical modes

Nonstationarity on Solar Cycle Timescales

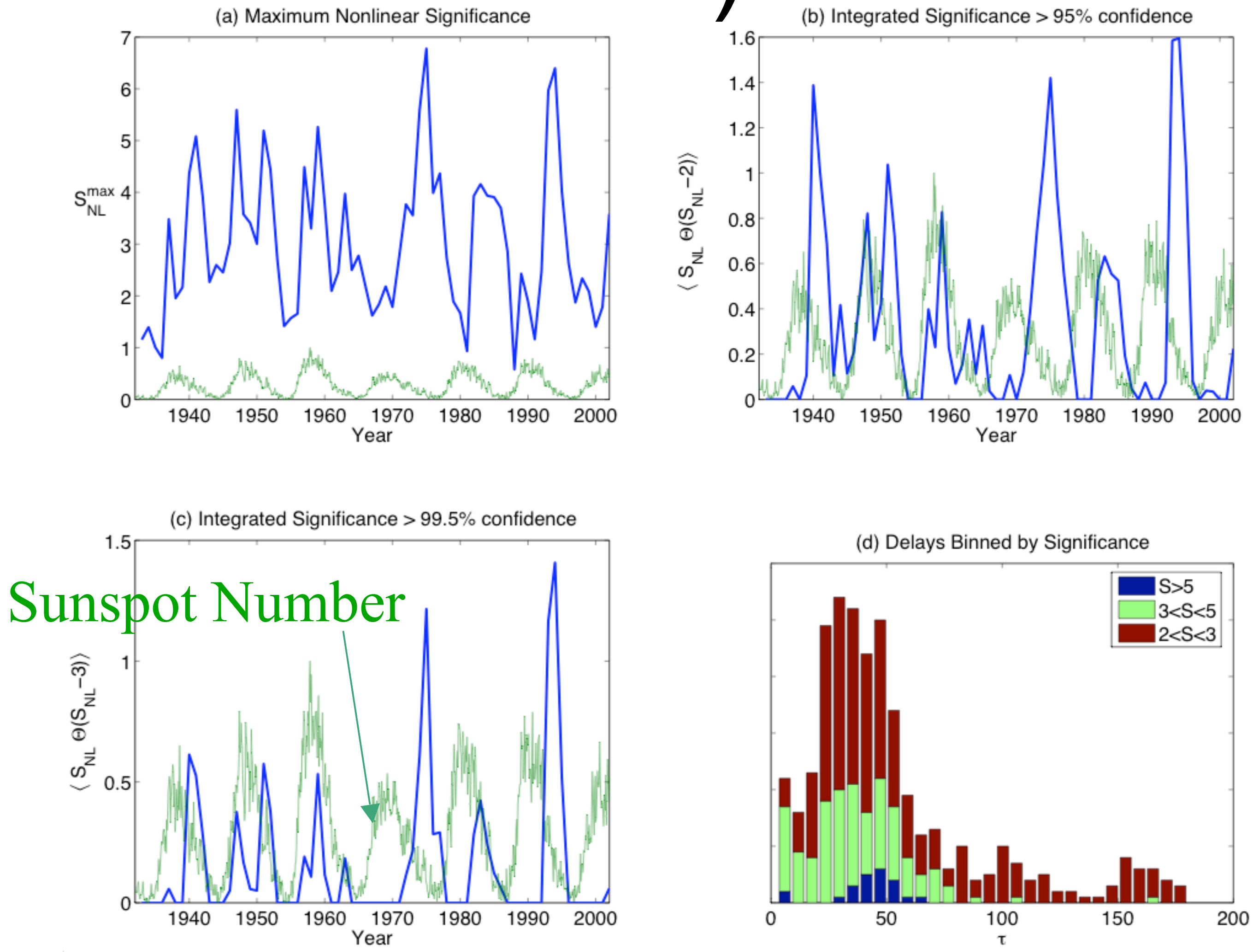
Example---Solar Maximum (MI)



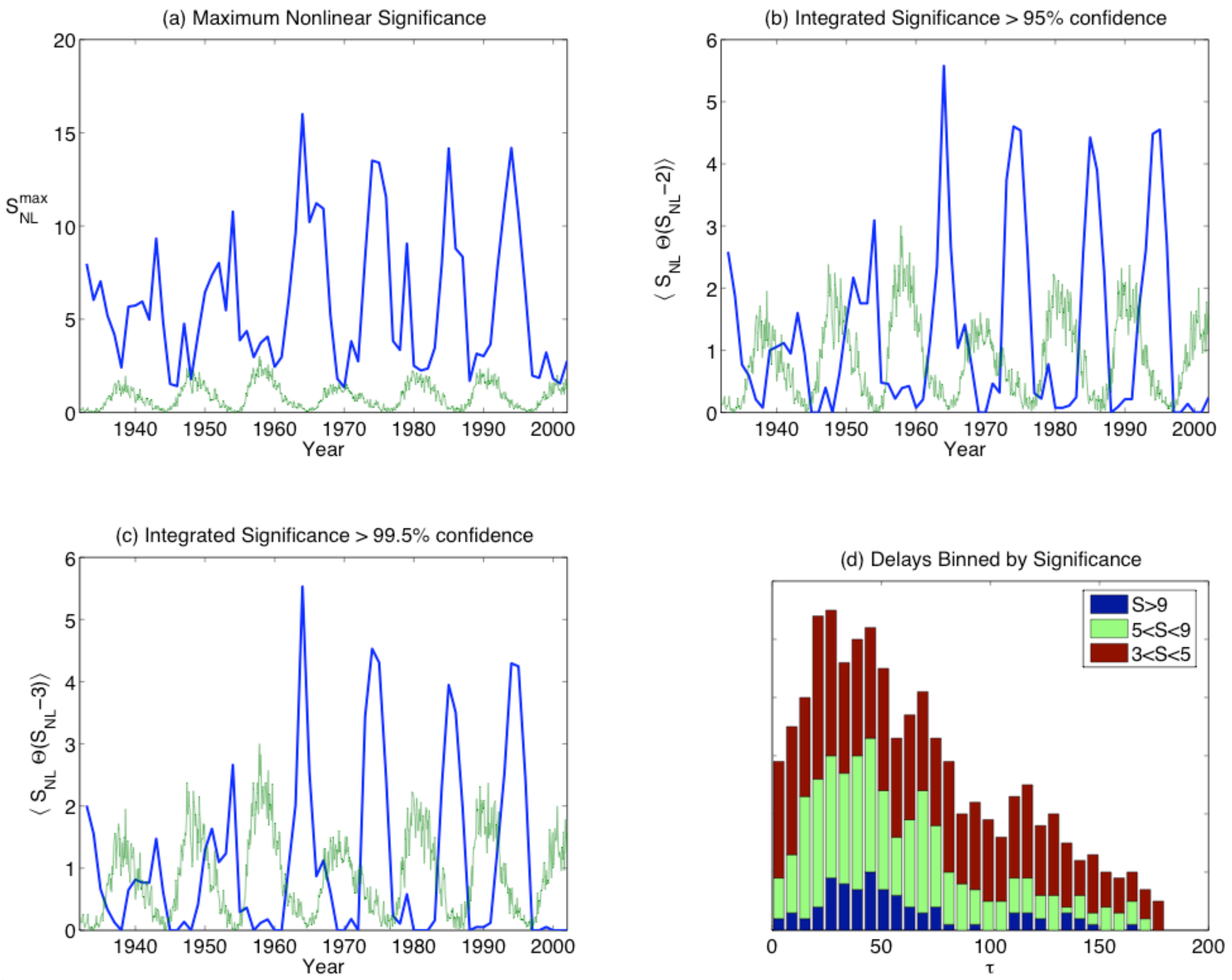
Example---Solar Minimum (MI)



Discriminating Statistic---(Λ - λ)



Discriminating Statistic---(C)



(a) Maximum Nonlinear Significance

(b) Integrated Significance > 95% confidence

(c) Integrated Significance > 99.5% confidence

(d) Delays Binned by Significance

Results of the Analysis

- Identified nonlinearity in magnetospheric dynamics during the declining phase of the solar cycle
- Identified a timescale (information horizon)
- The nonlinear response is an internal magnetospheric response to solar wind velocity enhancements
- Suggests an important nonlinear coupling sensitive to solar wind velocity

Planned Activities

- Build a database of direct measures of magnetospheric state
- Develop MI/Cumulant analysis to
  - characterize the underlying dynamics
  - discover the most important nonlinearities
  - determine information horizon
  - obtain a coupling function

  - investigate dimensionality
  - compress data stream through dimensional reduction